

Napsať somorôžkovi mail, hradby \rightarrow boxplotu nie sú $\pm 1,5 \text{ IQR}$
 Ťa učen dátá + zbielení

D.ô. \rightarrow inštalovať.

R + edition

? koniec štatist
 \rightarrow X-kvadrát
 $\text{? det} = \text{Multi. číslo}$

Počítačová štatistika

1976 - Bell labs S - štatistcs (náročka na C)

• attach: kópic stlpecov v glob. namespace • Ctrl+R - run

• logické: &, |, == != !, ! • Vektory = stlpcy

• order: rektor pozícií v pôv. • Násobenie matíc % %

• rank: pozicie v usp. poli • solve(A, v) \rightarrow

• data.frame: matice objektov riešenie $A \vec{x} = \vec{v}$

• objekt \$ param - stlpec

1988 - S+ - konservívna implementácia S

→ 1994 - R - Robert Gentleman, Ross Ihaka, odlišnosť od S

SAS - konservívny super R



NaN - net a number

• NA - not available = chýbajúci dátum

• imputačná technika - pokus nahradiť NA nejedýnnimi dátami

• FACTORS - premenne, ktoré nie sú číselni, LEVELS OF FACTOR - hodnoty tejto premennej (set: MF)
 table → kontingenčná tabuľka

zob

pre ktoré je záťaž med →

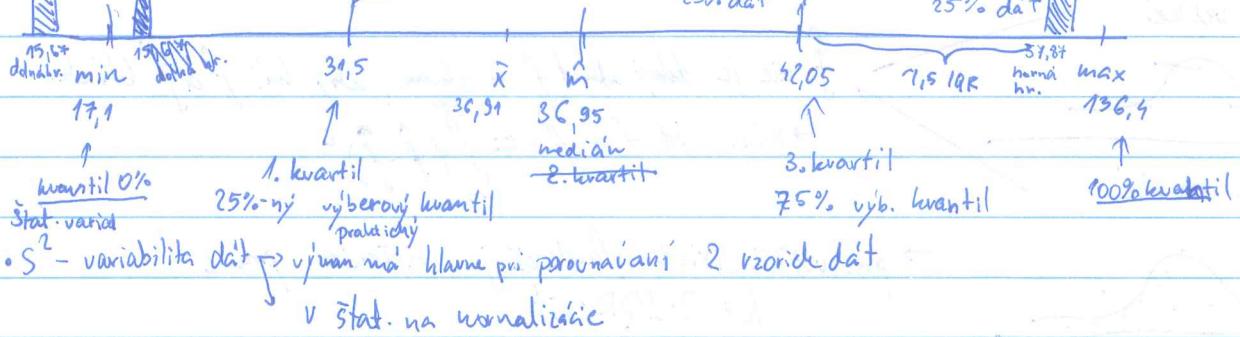
Bonusová úloha: $Y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$ $X = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$ $\left\{ \begin{array}{l} \text{one liner, bez for} \\ \text{n nepoznáme } n = \text{length}(Y) \\ \text{len príklady z 1.r, 2.r} \end{array} \right. \rightarrow \text{poslat e-mailom}$

$$\text{med} \quad \left\{ \begin{array}{l} \text{SIEGEL'S ESTIMATOR of intercept} \\ \text{v lin. regresii} \\ Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \end{array} \right.$$

$$X = \begin{pmatrix} x_1 & \dots & x_n \end{pmatrix} \left\{ \begin{array}{l} x_j * y_i - x_i * y_j \\ x_j - x_i \end{array} \right\}$$

Statistiky popisly a variability

\hat{m} → median, \bar{x} - priemer



• S^2 - variabilita dát \rightarrow významná hlavne pri porovnaní 2 rôznych dát

v štat. na normalizáciu

• IQR - interquartile range - uchopiteľnejšie ako S^2

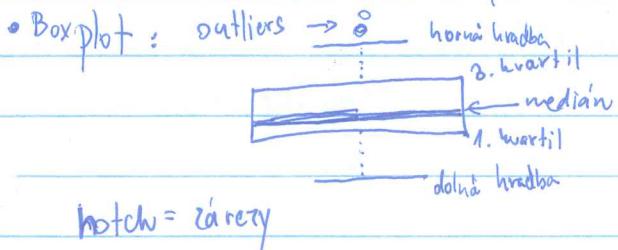
• dolná hradba = 1. kvartil - $1,5 \times \text{IQR}$: prečo 15^2 - určuje si hradieb - pre bežné dátá funguje

• outliers \rightarrow mimo hradieb \rightarrow treba sa pozrieť na to, ako vznikol? \rightarrow zazadit

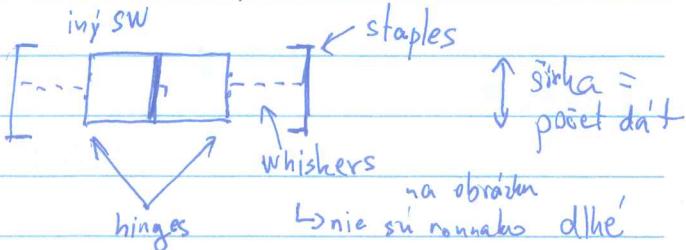
• outliers \rightarrow mimo hradieb \rightarrow treba sa pozrieť na to, ako vznikol? \rightarrow vylučiť (čísla výnimky)

Obrásky

Lotérija: keď si stavil číslo 0-999, tiež si tipli vyzbrované si rozdelili počiare



Notch = ráreky

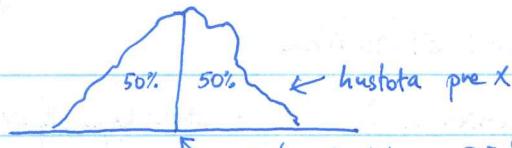


2 propagátor mediana? Schuster málo te FDI
2 preto trvá kresťanie tak isto? veta bodov

- analýza čas. radov
- robustné metódy

Zábery

X - náh. premena



m median husty s $P = \frac{1}{2}$ bude výhra > m

odhad pre m

1) bodový odhad: \hat{m} ... median z dát (výberový)



2) intervalový odhad: IS pre m: $\hat{m} \pm \text{čosi}$

Boxploty sa väčšinou malujú viacere'

Median je robustný \Rightarrow odľahlosť voči outlierom $5, 1, 1, 2, 1 \xrightarrow{\phi=2} \hat{m}=1$

median + priemer: Viac ako polovica Slovákov má podpriem. mzdú

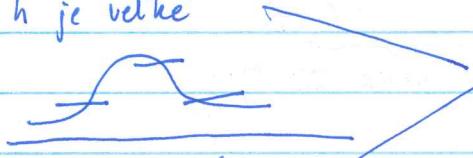
$$\min. mzd = 405 \quad \hat{m} = 703 \quad \bar{m} = 888$$

Norm. rozdelenie $N(0,1)$ $f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$

Cauchyho rozdelenie $f(x) = \frac{1}{\pi(1+x^2)}$

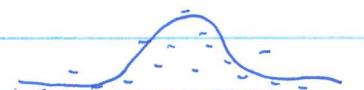
Freedman-Diaconis: šírka stĺpča histogramu (h)

a) h je veľké



\hat{f} nie je dobrý odhad f - chce, aby boli f a \hat{f} blízko
 \Leftrightarrow min. vzd f a \hat{f} $= \int_{-\infty}^{\infty} (f - \hat{f})^2$

b) h je malé

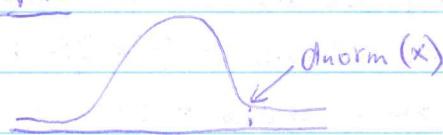


\rightarrow pre veľkú triedu funkcií je riešením práve
 $h = 2 \cdot \text{IQR} \cdot n^{-\frac{1}{3}}$

Rozdelenia a hustoty v R

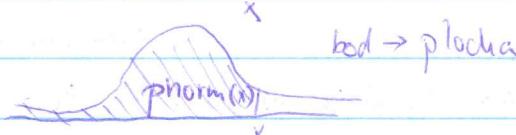
dnorm(x, mean=4, sd=3)
density norm. rozd

σ^2
hustota $N(4, 9)$



pnorm(7) = distrib. f. v 7 = $F(7) = P[X < 7]$

prob.
pnorm(0.09) = 9% kvantil



quantile



qnorm(0.95)

\rightarrow vytvára sa $X_1, \dots, X_{103} \sim N(0,1)$ a sú IID

* mnoho štat. metod využíva

čočsie generovanie dát

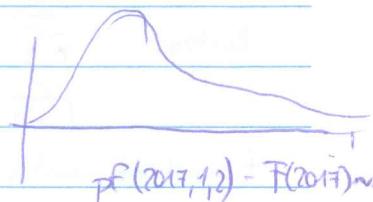
-norm - normálne rozdelenie

-t(..., df) - Studentovo rozdelenie t_{df}, df = degrees of freedom

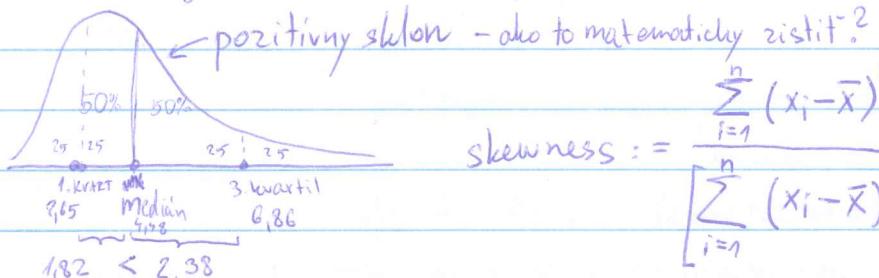
- chisq(..., df=4) - χ^2 - chi kвадрат

- f(..., df1=2, df2=8) - Fischerovo-Schnedekerovo rozdelenie

Sklon dát



Väčšina dát je z norm. rozdelenia, ale treba to overiť!



$$\text{skewness} := \frac{\sum_{i=1}^n (x_i - \bar{x})^3}{\left[\sum_{i=1}^n (x_i - \bar{x})^2 \right]^{\frac{3}{2}}} \rightarrow$$

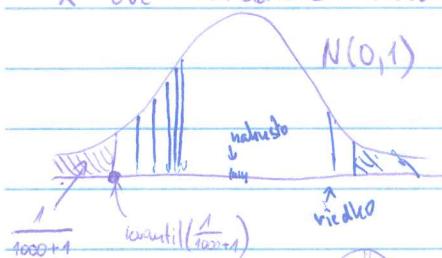
> 0 - pozitívny sklon
 $= 0$ - symetrická hust.
 < 0 - negatívny sklon

pozitívny sklon: $(3\hat{k} - \hat{m}) > (\hat{m} - 1\hat{k})$ (jednoduchosť, ale dátu treba usporiadaj)

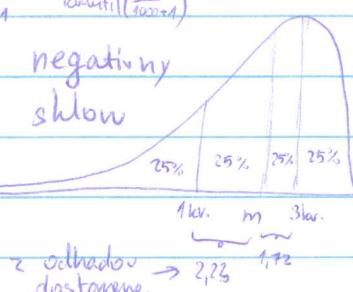
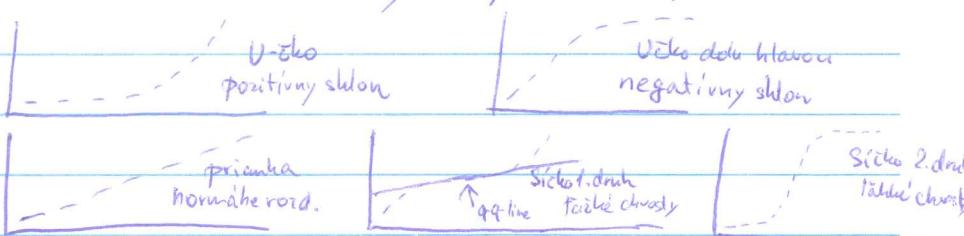
Q-Q plot (Quantile - Quantile plot - kvantilový diagram)

y-ové súradnice sú usporadane dátu $x_{(1)} < x_{(2)} < \dots < x_{(1000)}$

x-ové súradnice kvantily $N(0, 1)$ $\text{kvantil}(\frac{1}{1000+1}) < \text{kvantil}(\frac{2}{1000+1}) < \dots < \text{kvantil}(\frac{1000}{1000+1})$ (qq norm)



Môže nadebiadať rôzne tvary - vysvetliť tvary



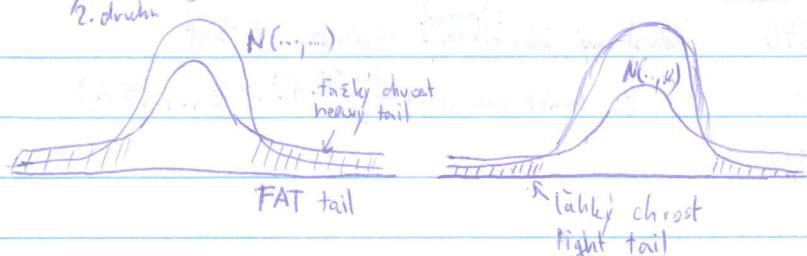
Ukazovatele sklonu (skewness, \hat{m} , 1.kv., 3.kv.)

niekedy nestačia \hat{m} , $\hat{m}/2$, $\hat{m}/3$

Síčkový kv. diagram: Dátu pochádzajú z hustoty, ktorá má ľahšie chvosty než $N(\dots, \dots)$, 1. druhu a teda špicatejší kopček než $N(\dots, \dots)$

qqline - priamka cez body $\frac{1}{4}, \frac{3}{4}$

Síčkový kv. diagram: Dátu sú z rozdelenie s hustotou s ľahšimi chvostami a mohutnejším kopčekom



Simulačné metódy - generovanie náh. dát a testovanie náhodnosti

KURTOSIS (vydutosť) - výberaj koficient spicatosti

$$\text{Kurtosis} := \frac{\sum_{i=1}^n (x_i - \bar{x})^4}{\left[\sum_{i=1}^n (x_i - \bar{x})^2 \right]^2}$$

≥ 3 spicajšia húpka, fážsie chvosty
 $= 3$ dátá sú z norm. rozdelenia
 < 3 mohutnejší húpok, ľahšie chvosty

teoretická kurtosis - získané ju z hustoty $= \frac{E((X-E(X))^4)}{E((X-E(X))^2)^2} = 3 \rightarrow$ bez ohľadu na μ a σ^2

$E((X-E(X))^2) = \sigma^2$
 $\underbrace{\sigma^4}_{\text{ešte treba dodátať že číslat je } 3\sigma^4}$

Reálne dátá - hľadáme hustotu, ktorá sa najviac podobá na naše dátá

- metóda: histogram (schodikový odhad)

- metóda jadrovej odhadu (kernel estimates) - v Rku density

Obrazok - histogram, gauss, kernel estimate \rightarrow ak sa podobá kernel est na gaussa
pravd. to bude norm. rozdelenie

Historické príklady falošania dát

- franc. profesor - hodí mincami 1000 krát a odvzdujte zápis \rightarrow ti čo podvádzali nemali dôležité rovnaké schr.

DV: kolko sérií bude mať dátum 8/X

- Mendel (Brnecký mincič, objavil zákon dedičnosti) - križenie hrachov s bielym zeleným kvetom

- porad sa na dátu Fischer (štatistik)

\hookrightarrow zákon siče platí, ale experimenty (veľa hrachov)

sú sfalošované

B	+	(2)
recessivny male	↓	domin. znak
gen. F1	100% (2)	
gen. F2	(3) ↓	(2)

25% ; 75%

100 hrachov na 1 poličku: 26:74, 25:75, 24:76

\hookrightarrow skutočné pomery by boli ďalej od 25:75 \rightarrow ukáza pomocou testu dobréj zhody

pravdep. za to môže pomocník: zaznal výsledky 85:15 \rightarrow CONFIRMATION BIAS

zberanie len dát, ktoré potvrdzujú hypotézu

Štatistické testy

1878: verilo sa, že rýchlosť svetla je 299 840, Michelsonov $\bar{X} = 299 852,6$

$$X_1, \dots, X_n \sim N(\mu, \sigma^2)$$

\hookrightarrow Studentov t-test

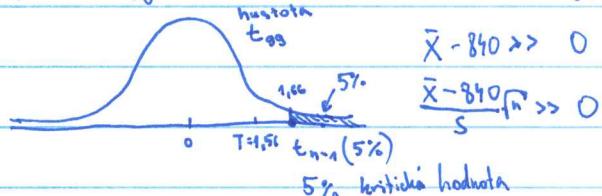
neponívame, skutočná rýchlosť svetla

$$H_0: \mu \leq 299 840$$

$$H_1: \mu > 299 840 \text{ (research hypothesis)}$$

One-sided t-test

MZR: H_0 zamietame ak $\bar{X} > 299 840$ Jednostranný, jednosmerový studentov t-test



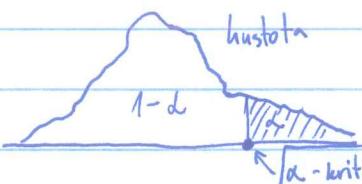
$$H_0 \text{ zamietame ak } T = \frac{\bar{X} - 299 840}{\sigma / \sqrt{n}} > t_{n-1}(5\%)$$

t-test v Rku t.test(y, alternative=greater, mu=840)

$$t = T = \frac{\bar{x} - 840}{S} f_n = 1,5694$$

testová statistika

chyba tu kritická hodnota $t_{gg}(5\%) \rightarrow$ Rku nenie rátaný limit hodnoty ale kvantily



"guličkový postup"

α -kritická hodnota = 1-d kvantil

$$T \neq t_{gg}(5\%) \quad t_{gg}(5\%) = 1,66$$

↪ Hypotezu H_0 neramietame

- Michelson tvrdí, že $c > 299840$, Studentov t-test zohľadnil

1) rozdiel 840 a 852 je dostatok malý

2) morka 100 dát je malá

3) dátá sú rovnomerné

$$p\text{-value} \in (0,1)$$

$$p = 5,98\%$$

↪ plocha od $T \rightarrow \infty$

Ak $p\text{-value} < 5\%$, H_0 zameňte

$p\text{-value} > 5\%$, H_0 neramietame

(G) "guličkový postup" porovnaj t a T

"p-value" porovnaj plochy 5% a $\int_{-\infty}^T p(x)dx$ ak $p\text{-value} = 5\% \rightarrow$ poslat dátu Somoržíkovi

Two-sided t-test.

TRUE: 299792 km/s

Porovnajme Michelsonove dátu s realitu $H_0: \mu = 792$ $H_1: \mu \neq 792$

MZR: H_0 zameňte ak $\bar{x} > 792$ alebo $\bar{x} < 792$

$$\bar{x} - 792 > 0 \quad \text{alebo} \quad < 0$$

$$\frac{\bar{x} - 792}{S} f_n > 0 \quad \text{alebo} \quad < 0$$

$$T = \frac{\bar{x} - 792}{S} f_n > t(2,5\%) \quad \text{alebo} \quad < -t(2,5\%)$$

$$7,6445 > 1,98$$

$$-1,98$$

p-value: plocha $T \rightarrow \infty$, $-T \rightarrow -\infty$

↪ zameňte H_0 , problém! - systematická chyba merača,

Ešte jeden 1-stranný t-test Michelson dostával obita s $\mu > c$

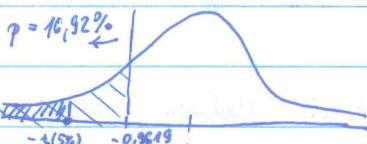
V 1877 sa verilo, že NEWCOMBE: 860 (MICHELSON: 852,4 → bol bližšie k TRUE: 792)

$$H_0: \mu \geq 860 \quad \text{vs.} \quad H_1: \mu < 860$$

H_0 zameňte ak $\bar{x} < 860$

$$T = \frac{\bar{x} - 860}{S} f_n < -t(5\%) = t(95\%)$$

$$-0,9619 > -1,66 \Rightarrow H_0 \text{ neramietame}$$



test nie je štatisticky významný

$$p\text{-value} = p(T < -0,96) = F_T(-0,96)$$

* JOHN TUKEY - vynálezca box plotu 1961
ovz. bit (1948)
software (1958)

$$X_1, \dots, X_n \sim N(\mu_x, \sigma^2) \quad \text{predpokladané,}\braket{\text{je sú rovnaké,}} \\ Y_1, \dots, Y_n \sim N(\mu_y, \sigma^2) \quad \text{treba overiť} \quad H_0: \mu_x = \mu_y \quad \text{vs.} \quad H_1: \mu_x \neq \mu_y$$

skutočné rýchlosť súčtu v určitých dňoch

Predtest $H_0: \sigma_x^2 = \sigma_y^2 \quad \text{vs.} \quad H_1: \sigma_x^2 \neq \sigma_y^2$

MZR: H_0 zamietneme ak $S_x^2 > S_y^2$ alebo $S_x^2 < S_y^2$

F test $\frac{S_x^2}{S_y^2} >> 1$ alebo $\frac{S_x^2}{S_y^2} << 1$ (var. test)

2. a 4. dň $\frac{S_x^2}{S_y^2} = F = 1,0344 \quad p\text{-value} = 93\% > 5\% \Rightarrow H_0$ neramietane 😊

2-výberový Studentov t-test

H_0 ~~zamietame~~ $\mu_x > \mu_y$ alebo $\mu_x < \mu_y$

var.equal = TRUE \rightarrow predtest dopadol dobre

p-value = 7,17% > 5%, teda H_0 neramietane

\hookrightarrow t-test prevedal, rozdiel ($\bar{x} = 856$, $\bar{y} = 820,5$) nie je štatisticky významný

lebo morka 20 dát je málo

1. a 2. dň ~~$H_0: \sigma_x^2 = \sigma_y^2$~~ vs. $H_1: \sigma_x^2 \neq \sigma_y^2 \quad \frac{S_x^2}{S_y^2} = 2,94$ 😕

\hookrightarrow 2-výberový Studentov Welchov t-test var.equal = FALSE

1, $T = \frac{\bar{x} - \bar{y}}{\sqrt{s_x^2 + s_y^2}} \sim t_{\nu}$ (tento test je približný)

2, $\nu = ?$, treba ho odhadnúť (#st. volnosť)

p-value = 6%, $H_0: \sigma_x^2 = \sigma_y^2$ neramietame (dát je málo a majú veľkú varianciu)

Párový t-test

Predpoklad sl. t-testu a Welchovo t-testu: dátá X_i a Y_i sú neráviasle'

Dátá podriaďole: opotrebovanie pred a po nosení, rozdiel je 1 číslo

- t diela nosilo topinky A aj B, porovnáme opotrebu material A: X material B: Y

$H_0: \mu_x \leq \mu_y \quad \text{vs.} \quad H_1: \mu_x > \mu_y$ (research hypothesis)

A sa opotrebova viac ako B.

Nemôžeme spraviť studentov, welchov t-test, lebo dátá ^{nie sú} neráviasle' (v stĺpcach, pravá / ľava, topinky)

\hookrightarrow použijeme párový test

\rightarrow dátá v dvojiciach $(X_1, Y_1), \dots, (X_n, Y_n) \sim N_2 \left(\begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}, \begin{pmatrix} \sigma_x^2 & \text{cov} \\ \text{cov} & \sigma_y^2 \end{pmatrix} \right)$

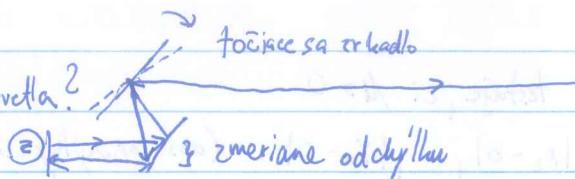
\rightarrow ① párovomie $Z_i = X_i - Y_i \sim N(\underbrace{\mu_x - \mu_y}_{\mu}, \underbrace{\frac{\sigma_x^2 + \sigma_y^2 - 2\text{cov}}{n-1}}_{\sigma_Z^2})$

stat. overiť, $H_0: \mu \leq 0 \quad \text{vs.} \quad H_1: \mu > 0$

je rozdiely \rightarrow ② 1-súborový Studentov t-test na dátach Z_1, \dots, Z_n ($t = \frac{\bar{Z} - 0}{\sqrt{s_Z^2}}$ t_{n-1} (%)

Ako do t-testu dám párové dátá, bude sa správať konzervatívne, vysoká p hodnota

Ako Michelson meral rýchlosť svetla?



problemy: kde svetlo prešlo tam a naspäť, orhadlo sa v ľavej nestihlo pohybiť

- (poliat) → rýchlosť potreboval porozvíti orhadlo → použil parný stroj = zdroj nepresnosti (nestabilnosť)
- prvý Američan, ktorý získal nobelovu cenu za vedu

Testy normality

Kolmogorov-Smirnov test

$$X_1, \dots, X_n \stackrel{\text{?}}{\sim} N(\mu, \sigma^2) \quad H_0: \text{dáta } \sim N(\mu, \sigma^2) \text{ vs. } H_1: \text{dáta nie sú } \sim N(\mu, \sigma^2)$$

dáta pochádzajú z rozdelenia, ktoré má distribučnú funkciu $F(\cdot) = ?$ (neznáme ju)

CDF: Cumulative Distribution Function

$$\hookrightarrow \text{odhad } F(\cdot) = P(X < \cdot) = \frac{\#\{x_i < \cdot\}}{n} =: \hat{F}(\cdot) \quad \text{--- ECDF: Empirical CDF}$$

Idea testu Ak plati H_0 , tak $\hat{F}(\cdot)$ by sa mala podobáť na CDF pre $N(\mu, \sigma^2)$
za μ zvolime \bar{x} , za σ^2 zvolime S^2

H_0 zamietane ak $\hat{F} \neq \text{pnorm}(\bar{x}, S)$ sú veľmi odlišné

D: maximálna zvislá odchyľka $\Rightarrow D$ (test. štatistika)

pre rýchlosť svetla $p \approx 45\% > 5\%$, H_0 neramietane

p hodnota: D sa riadi nejakyim rozdelenim
Kolmogorov-Smirnov ho značili

Shapiro-Wilkov test (pre male sady dát)

Wilk ≠ sir Samuel Wilkes

- snaziť sa zistiť, či kuantilový diagram vyzera ako priama

Záden test normalitu Z ; (podrážky: dátá A - dátá B) neramietol napriek tomu,
že histogram naznačuje, že dátá sú vlastne normálne, ale vzhľad z 10 dát

Neparametrické testy

Levi-Strauss - testovali: novaz namáhašanie farby ťažmi vs. strojmi na látke

- dátá: % nepodarok ťažia - stroje \rightarrow o kolko % sa ťažia viac než stroje

\rightarrow dát je len 22, ale napriek tomu SW test normalitu ramietka

\rightarrow je tu veľa outlierov, pri normálnom rozdelení by boli byť približne $\frac{1}{100}$ dát

$$X_1, \dots, X_n \sim \mathcal{N}(\mu, \sigma^2)$$

? $\mu = E(X)$ - stredná hodnota nepodarok ťaži voči strojom

$$H_0: \mu \leq 0 \text{ vs. } H_1: \mu > 0 \quad (H_1: \text{ ťažia sú horši ako stroje})$$

- numeziame použiť studentov t-test (nemáme N)

Wilcoxon (1945) : testuje, ci $\mu > 0$

1) Zábernice $|x_1 - 0|, |x_2 - 0|, \dots, |x_n - 0|$ (abs. odchýlky od strednej)

2) Oranujene odchýly $x_{(1)}, \dots, x_{(n)}$... 1 ... 3 ... $x_{(n)}$ na tejto pozícii je najväčšia = R_1, \dots, R_n

3) $S_W = \sum_{i, x_i > 0} R_i$

(súčet rankov, kde bola odchýlka kladná) H_0 zameňame ale $S_W > 0$

$X_1, \dots, X_m \sim N(\mu_x)$

$H_0: \mu_x = \mu_y$ vs. $H_1: \mu_x \neq \mu_y$

$Y_1, \dots, Y_n \sim N(\mu_y)$

sú navzájom nerelativné

1) Jedna sada dát: $X_1, \dots, X_m, Y_1, \dots, Y_n$

2) Ranky: $1, \dots, m, \dots, n = R_1, \dots, R_m, R_{m+1}, \dots, R_{m+n}$

3) $S_W = \sum_{i=1}^m R_i$ - súčet len rankov X ov

- potreba overovať rovnosť disperzii

- Ak do t-testu vložíme neuzávislé dátá (majú veľa outlierov), t-test sa správa konzervatívne

Ak test $\sigma_1^2 = \sigma_2^2$ dopadne tesne k 5% \rightarrow skúšime Welch \rightarrow nebude fungovať

B-D-U. o 2 roky po Wilcoxonovi: MANN, WHITNEY 1947 vymysleli ďalší jednoduchší postup

vzťahy dvojice $X_i \leq Y_j$, Test. štat. $S_{MW} := \#\text{prípadov } \{X_i > Y_j\}$

- je ich $m \cdot n$

- Wilcoxonova test. štatistika - MANN, WHITNEY = konšt.

1) Zistiť, či sú dané konstanta rôzne

2) Dôkázať, že možnosť bude rovnaká pre všetky prípady

štat. nie je závislý od n, m , nebude závisieť od konkrétnych dát

Sestry a ruhovice - používajú ich zdravotníctvo? tajnú ich sledovali \leftarrow pri práci s klientom

školenie \rightarrow potom nové priestúpky o 1 mesiac

o 2 mesiace \rightarrow horšie

o 5 mesiacov \rightarrow aké helby školenie aby nebolo

Bernoulliho schéma - o palujúce statek tento istý experiment s pravdep. $p \rightarrow$ binomické rozd.

$X \sim \text{Bin}(n, p)$ n - # sledovaní urámcu obdobia

\rightarrow kolikačt ich použili? p - svedomitosť cestier - nepoznáme (všetkých / priemerne) 8

$$X = \sum \text{kedy vysili} \quad \hat{p} = \frac{x}{n}$$

$n = \# \text{pozorovani}$

1-strany test:

v NEW YORKU JE PRIEM. SVEDOMITOSŤ 20 %

$$H_0: p \leq 20 \quad \text{vs. } H_1: p > 20$$

MZR: H_0 zamietane ak $\hat{p} \gg 0,2$

$$\hat{p} - 0,2 \gg 0$$

$$X \sim \text{Bin}(n, p)$$

$$\textcircled{1} \text{ LAPLACE-MOIVRE CLT: } \frac{X - np}{\sqrt{np(1-p)}} \sim N(0,1)$$

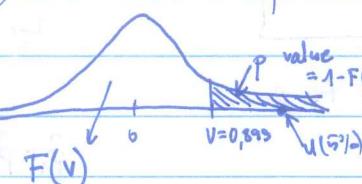
$$\textcircled{2} \text{ MMS: } \frac{1}{n} \left(\frac{X}{n} - p \right) \sim N(0,1)$$

← Dálej: akým rozdeleniu

sa riadi $\hat{p} - 0,2$?

$$\frac{\hat{p} - 0,2}{\sqrt{\frac{p(1-p)}{n}}} \Rightarrow u(5\%)$$

tedy H_0 zamietane



Mali sme tušenie, že priemer nasledu sestier

je väčší ako novopreky

Ale nie je to štat. významné:

rozdiel 20% a 25% je malý

$n = 51$ je málo dať

③ B.D.U.

treba dôkázať

$$\frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \sim N(0,1)$$

aj keď \hat{p} používané

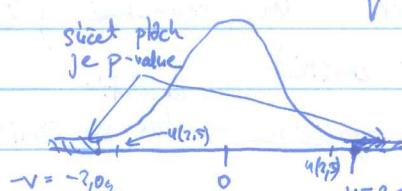
v odhadu disperzie

- treba použiť konverg. podľa distr. funkcie
- treba použiť kon. podľa pravdepodobnosti
- zákon velkých čísel
- Slutsky theorem

objektívny test:

v USA je priemer svedomitosti 13%

$$H_0: p = 0,13 \quad \text{vs. } H_1: p \neq 0,13$$



MZR: H_0 zamietane, ak $\hat{p} \gg 0,13$ ale $\hat{p} \ll 0,13$

H_0 zamietane, ale $\frac{\hat{p} - 0,13}{\sqrt{\frac{p(1-p)}{n}}} > u(2,5\%) \approx 2,5\%$

$$p\text{-value} = 4\% < 5\% \Rightarrow H_0 \text{ zamietane}$$

$$H_1: p \neq 0,13$$

Čo je to p-value?

Čo si ľudia myslia: Keď $p < 5\%$, tak H_0 zamietane. Ak $p > 5\%$ H_0 neramietane:

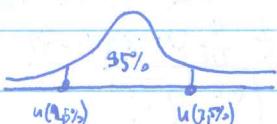
$$\Rightarrow \text{p-va} \cancel{\text{lue je }} P(H_0 \text{ platí})$$

(H_0 platí nás je výhodná udalosť)

extremnosťia

p-value: je pravdepodobnosť, že testovacia štatistika by býť ešte väčšia ako teraz aspon tak extrema

Interval spoločnosti



$$P\left(-u(2,5\%) < \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} < u(2,5\%)\right) = 95\%$$

$$P\left(\frac{\hat{p} - u(2,5\%)}{\sqrt{\frac{p(1-p)}{n}}} < p < \hat{p} + u(2,5\%) \frac{\sqrt{\frac{p(1-p)}{n}}}{\hat{p}}\right) = 95\%$$

$$L = 0,135$$

$$\hat{p} = 0,254$$

$$U = 0,374$$

Závisí svedomitosť od skúsenosti?

skutočná svedomitosť $\in (13\%, 37\%)$

stred intervalu

$$X_1 \sim \text{Bin}(n_1, \hat{p}_1) \quad \text{najviac 3 roky} \quad \hat{p}_1 = \frac{4}{10}$$

$$X_2 \sim \text{Bin}(n_2, \hat{p}_2) \quad \text{viac ako 3 roky} \quad \hat{p}_2 = \frac{6}{41}$$

Zberka: zvláštna príručka 2. kapitola
t-test pravdepodobnosť.

2-sample 2-sided test

$$H_0: p_1 = p_2 \text{ vs. } H_1: p_1 \neq p_2$$

MZR: H_0 zameňuje ak $\hat{p}_1 > \hat{p}_2$ alebo $\hat{p}_1 < \hat{p}_2$

H_0 zameňuje ak $\hat{p}_1 - \hat{p}_2 > 0$ alebo < 0

$$V > u(2,5\%) \text{ alebo } < -u(2,5\%)$$

pričom $p_1 - p_2 = 0$
predpokl.

$$V = \frac{\left(\hat{p}_1 - \hat{p}_2\right) - (p_1 - p_2)}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}} \sim N(0,1)$$

Dá sa odvodiť rovnaké ako 1-sample

Je štatisticky významné meranie svedomitosti
 \hookrightarrow (z 1-str. testu potom zistíme, že nie je skúšané
sú svedomitejšie)

$$\hat{p}_1 - \hat{p}_2 = 0,554 - \text{bodený odhad}$$

$$IS P(-u(2,5\%) < V < u(2,5\%)) = 95\%$$

$$P\left(\hat{p}_1 - \hat{p}_2 - \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} < p_1 - p_2 < \hat{p}_1 - \hat{p}_2 + u(2,5\%) \right) = 95\%$$

$$IS = (0,249, 0,1857) - možno sú len o 25\% svedomité$$

\hookrightarrow preto výsiel taký široký? : v prvej skupine (neskúšaných sestier) máme len 10 pozorovaní

$X \sim \text{Bin}(n, p)$ - je rozumné predpokladať, že každá sestra v nemocnici

\hookrightarrow možli by sme uvažovať p pre každú sestru zvlášť kde' rovnaké p ?
(logistická regresia - prilis zložité)

\hookrightarrow je to ale až také volouina? dôvoda prídužka \rightarrow dôv sa uniformizuje

Robert Box: "Every model is wrong, but some are useful."

zdrojová hustota
veličiny (X)

$$\begin{aligned} &\text{PEARSON CORRELATION COEF.} \quad \text{Korelačná analýza} \\ &S := \frac{\text{cov}(X, Y)}{\sqrt{D(X) D(Y)}} = \frac{E((X-E(X))(Y-E(Y)))}{\sqrt{D(X) D(Y)}} = \frac{\int \int (x-E(x))(y-E(y)) f(x,y) dx dy}{\int \int f(x,y) dx dy} = ? \end{aligned}$$

$$\text{dáta } (x_1, y_1), \dots, (x_n, y_n) \rightarrow \text{odhad pre } S: \hat{S} := \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{S_x^2 S_y^2}} \in [-1, 1]$$

$$\text{cor}(Income, Frost) = 0,22 \rightarrow \text{luboč bohatšie regióny sú na S/SV} - \text{nie je to kauzalita}$$

$$\text{cor}(Murder, Life.Ex) = -0,87 \rightarrow \text{nie, ie by sa urádzili} \Rightarrow \text{kraťte život / ale je tan eká situácia} \Rightarrow \text{kraťte život}$$

\hookrightarrow sú to len odhady \hat{S} - neviabí záveru len na tom \rightarrow TEST

Testy významnosti korelačie

$$H_0: S = 0 \quad \text{vs.} \quad H_1: S \neq 0$$

MZR: H_0 zamietane ak $\hat{S} > 0$ alebo $\hat{S} < 0$
 alternative = "two-sided", method = "pearson"

1) ideale testovanie, či sú dátá z norm. rozd \rightarrow ks-test

↪ population & area nie sú z N rozdelenia.

2) test. štatistika $\frac{\hat{S}}{\sqrt{1-\hat{S}^2}}$ porovnáme so t_{n-2}
 (test závisí aj od počtu dát)

cor (Income, Frost) \rightarrow p-val = 11% $\Rightarrow H_0$ nezamietane \rightarrow vymysleli sme ešte teóriu, ale nie je stáť významná.

$$\text{cor.test}(Income, HS.grad) \quad H_0: S \leq 0 \quad \text{vs.} \quad H_1: S > 0$$

MZR: H_0 zamietane ak $\hat{S} > 0$

\hookrightarrow p-val = 7e-7 $< 5\%$ $\Rightarrow H_0$ zamietane
 $X = \text{population}, Y = \text{frost}$ cor(pop, frost) = -0,33

$$H_0: S = 0 \quad \text{vs.} \quad H_1: S \neq 0$$

$\hookrightarrow X$ sa neviadajú normalným rozdelením



SPEARMAN corel. coef.:

1) dátá X_1, \dots, X_n Y_1, \dots, Y_n

ranky $X_{(1)}, \dots, X_{(n)}$ $Y_{(1)}, \dots, Y_{(n)}$

$$R_1, R_2, \dots, R_n \quad Q_1, Q_2, \dots, Q_3$$

$$2) \hat{S}_s := \text{pearsonovo } \hat{S} \text{ využitie z rankov} = \frac{\frac{1}{n} \sum_{i=1}^n (R_i - \bar{R})(Q_i - \bar{Q})}{\sqrt{S_R^2 + S_Q^2}}$$

$H_0: X$ a Y spolu negatívne vs. $H_1: X$ a Y spolu súvisia negatívne

MZR: H_0 zamietane ak $S_s < 0$

toto je konšt.

↪ \hat{S}_s je konšt.

↪ S_R a S_Q sú konšt.

↪ aj tota sú konšt.

p-value < 5% $\rightarrow H_0$ zamietane \rightarrow je stáť významná korelácia Income - Frost

$S_p = \hat{S} = 0,33 \neq \hat{S}_s = -0,46$ ale väčšinou majú rovnakú znamienku

* Pearson a Spearman - kolygovia na Cambridge, Pearson počítaval Spearmanom (nebol vyniesiel štatistik ale psycholog)

Spearman: Faktorová analýza: výsledok 10boja je určený len 2-3 faktormi: výtrvalosť, rýchlosť, sila

Fischerova Z-premenná

- IS pre S ale sú dátá z $N(\dots, \dots)$

$$(X) - \text{income} \quad H_0: S \leq 0 \quad \text{vs.} \quad H_1: S > 0 \quad \dots \text{či sú sú dátá s } S \text{ rovná?}$$

1) bodový odhad: $\hat{S} = 0,61$

↪ platí len odmiene pre $n \geq 3$
 pre $n > 3$ závisí od S :

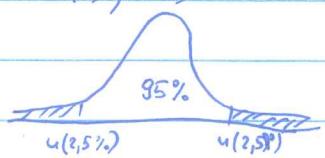
2) interval spoločnosti pre S \rightarrow oduvodí sa z rozdelenia $\hat{S} \sim N(S, \text{DIFERENCA})$

\hookrightarrow použijeme transformáciu: Fischer Z-transform = $\frac{1}{2} \ln \frac{1+\hat{S}}{1-\hat{S}}$ hyperbolický arctg.
 (Lahký dôkaz Taylorovým rozvojom)

$$\text{atanh}(\hat{S}) = \frac{1}{2} \ln \frac{1+\hat{S}}{1-\hat{S}} \sim N(\text{atanh}(S), \frac{1}{n-3})$$

$$95\% = P\left(-u(2,5\%) < \frac{Z - \text{atanh}(S)}{\sqrt{\frac{1}{n-3}}} < u(2,5\%)\right)$$

$$\frac{Z - \text{atanh}(S)}{\sqrt{\frac{1}{n-3}}} \sim N(0,1)$$



$$= P\left(Z - u(2,5\%) \sqrt{\frac{n-3}{1}} < \text{atanh}(S) < Z + u(2,5\%) \sqrt{\frac{n-3}{1}}\right)$$

$\rightarrow \tanh$ je rastúca f \rightarrow možno použiť

$$= P\left(\tanh\left(Z - u(2,5\%) \sqrt{\frac{n-3}{1}}\right) < S < \tanh\left(Z + u(2,5\%) \sqrt{\frac{n-3}{1}}\right)\right)$$

$$\text{IS} = (0,41, 0,76) \quad \text{nie je symetrický}$$

* je niečo ako Z-prem pre SPEARMANU $\sim N(\dots, \frac{1,06}{n-3})$, nie pre laboratórne rozdelenie

2-SAMPLE TEST

$(X) \sim \text{INCOME}$
 $(Y) \sim \text{HS. BRAD}$

$$\text{YUH: } \hat{S}_1 = 0,83 \\ S_1 = ?$$

$$\text{SEVER: } \hat{S}_2 = 0,20 \\ S_2 = ?$$

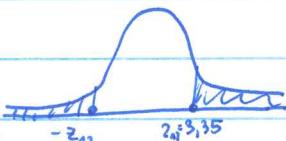
$$H_0: S_1 = S_2 \quad \text{vs.} \quad H_1: S_1 \neq S_2$$

$$Z_1 \sim N(\text{atanh}(S_1), \frac{1}{n_1-3}) \quad Z_2 \sim N(\text{atanh}(S_2), \frac{1}{n_2-3})$$

$$Z_1, Z_2 \text{ sú nezávislé} \Rightarrow Z_1 - Z_2 \sim N(\text{atanh}(S_1) - \text{atanh}(S_2), \frac{1}{n_1-3} + \frac{1}{n_2-3})$$

$$\boxed{\frac{(Z_1 - Z_2) - (\text{atanh}(S_1) - \text{atanh}(S_2))}{\sqrt{\frac{1}{n_1-3} + \frac{1}{n_2-3}}} \sim N(0,1)}$$

$$\Rightarrow \text{TESTOVÁ STATISTIKA} \quad \frac{Z_1 - Z_2}{\sqrt{\frac{1}{n_1-3} + \frac{1}{n_2-3}}} =: Z_{1,2} \quad \text{porovnáve s u (2,5%)}.$$



$$p\text{-val } 2(1 - \text{pnorm}(Z_{1,2})) < 5\%$$

TEÓRIA { SEVER: aj nezdaní si vedia zarobiť - lepsi systém $\rightarrow H_0$ zamietane
YUH: nezdaní sú kopači

Lineárna regresia

$$Y_i = \underbrace{\beta_0}_{\text{parametrické regresie}} + \underbrace{\beta_1 x_i}_{\text{teplota}} + \underbrace{\varepsilon_i}_{\text{chyba regres. modelu}}, \quad i=1 \dots n \quad \text{číslo dát}$$

$$\begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} = \underbrace{\begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}}_{\mathbf{Y}} \underbrace{\begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}}_{\boldsymbol{\beta}} + \underbrace{\begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}}_{\boldsymbol{\varepsilon}}$$

Chceme zistit $\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} \stackrel{?}{=} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} = 2 \rightarrow$ treba ich odhadnúť

LS-estimator - LEAST SQUARES

zvisle

$$\hat{\beta}_0 = (-2,22) = \hat{\beta} \\ \hat{\beta}_1 = 0,07$$

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}$$

\downarrow súčet odchýlok = 0
 \uparrow súčet štvorcov

odchýlky = rezidua'

konst., ktorú si budeme voliť

$$\text{KONTRAST: } a_0 \beta_0 + a_1 \beta_1 = (a_0, a_1) \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} \stackrel{?}{=} a^T \boldsymbol{\beta} = 2$$

ODHAD PRE KONTRAST $a^T \hat{\boldsymbol{\beta}}$

PLATÍ $\frac{a^T \hat{\boldsymbol{\beta}} - a^T \boldsymbol{\beta}}{\sqrt{a^T (\mathbf{X}^T \mathbf{X})^{-1} a}} \sim t_{n-2}$

\uparrow v modeli sú 2 parametrické β

$$S^2 = \frac{\sum \varepsilon_i^2}{n-2} \leftarrow \sum_{i=1}^n (\text{reziduum})^2 = \sum_{i=1}^n \varepsilon_i^2 \quad \text{SUM of Squares}$$

parametrov β

\hookrightarrow zdroj dostaneame MMS IS pre β

$$IS \text{ pre } \alpha^T \hat{\beta} : ① \hat{\beta} \pm t_{n-2}(2,5\%) \cdot S \sqrt{\alpha^T X^T X \alpha}$$

ak robime \leftarrow 95% spol. je len pre 1 bod!

(kontrast)

viac odhadov
spolahlivosť hiesá

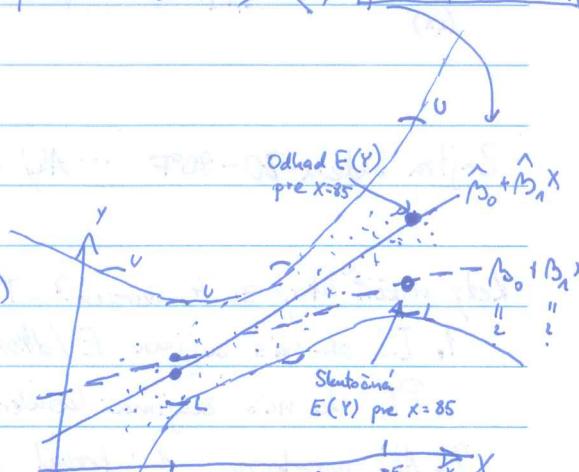
~~95%-nyj~~ pás spolahlivosti

$$\text{reziduum} = \begin{pmatrix} Y_1 - (\hat{\beta}_0 + \hat{\beta}_1 x_1) \\ \vdots \\ Y_n - (\hat{\beta}_0 + \hat{\beta}_1 x_n) \end{pmatrix} = Y - \hat{X} \hat{\beta}$$

$\hat{\beta}_1 = ?$ nepozname

1) Bodový odhad $\hat{\beta}_1 = 0,07$ (lin. regr.)

2) $(0,1) \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} = \hat{\beta}_1 \rightarrow$ do IS pre kontrast vložíme vektor $(0,1)$



$$x = 85^\circ F (29,4^\circ C) \Rightarrow E(Y) = ?$$

$$E(Y) = E(\hat{\beta}_0 + \hat{\beta}_1 \cdot 85 + \varepsilon) = \hat{\beta}_0 + \hat{\beta}_1 \cdot 85 + \overbrace{E(\varepsilon)}^0 = ?$$

1) Bodový odhad dosadiame za $\hat{\beta}_0, \hat{\beta}_1$ odhady $\hat{\beta}_0, \hat{\beta}_1 \Rightarrow \hat{\beta}_0 + \hat{\beta}_1 \cdot 85 = (1,85) \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} = (1,85)$ kontrast = $(1,85)$

2) Intervalový odhad pre $E(Y) : (3,61, 3,89) = (L, U)$

E - str. hodnota - zo základu všetkých čísel - ak veta k rátat zopakujeme pokus, $\bar{Y} \rightarrow E(Y)$

1953, Američan

Scheffého simultánne intervaly spolahlivosti pre $\alpha^T \hat{\beta}$

$$② \hat{\beta} \pm \sqrt{2F_{2,n-2}(5\%) S \sqrt{\alpha^T (X^T X)^{-1} \alpha}} \quad \begin{matrix} \text{Fischer-Schödlerov} \\ \text{určujú} \end{matrix} \quad - 95\% \text{ pás spolahlivosti}$$

\star ľahko dôkazat

parametrov Scheffého korelacia

Rozšírenie na okrajoch: v strede je veľa dát, na okrajoch je menej \rightarrow menší istoty odhad

Predikčný interval

Zajtra bude $85^\circ F \dots$ aký bude ozín? $Y = ? = \hat{\beta}_0 + \hat{\beta}_1 \cdot 85 + \varepsilon$

1) Bodový odhad $\hat{\beta}_0 + \hat{\beta}_1 \cdot 85 + 0$

2) Intervalový odhad = PREDIKČNÝ INTERVAL $③ \hat{\beta} \pm t_{n-2}(2,5\%) \sqrt{1 + \alpha^T (X^T X)^{-1} \alpha}$

zdroj na ekonometrii

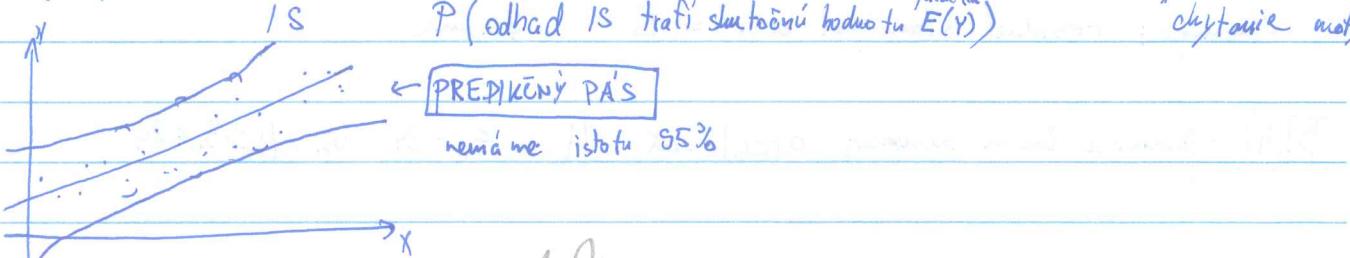
$\hat{\beta}_0 + \hat{\beta}_1 \cdot 85 + \varepsilon \leftarrow 3$ zdroje náštory

pozn) PI pre $Y : (2,58; 4,92)$ } $\frac{\text{Pretože PI}}{\text{Súčasne}}$ - suchý dôvod: keďže vo vzorci je 1+
IS pre $E(Y) : (3,61; 3,89)$ } $\frac{\text{Súčasne}}{\text{berie do úvahy } \varepsilon}$

$\hat{\beta}_0 + \hat{\beta}_1 \cdot 85 \leftarrow 2$ zdroje náštory

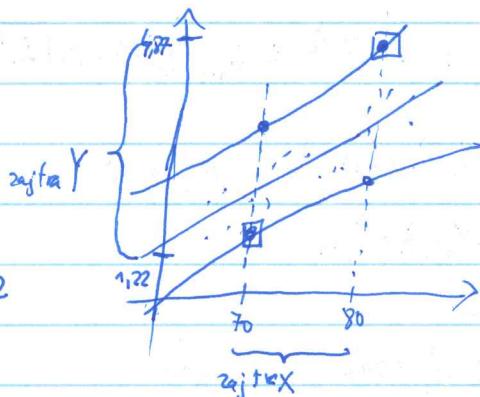
poz 3) Čo je predikčný interval? $P(\text{teplota zajtra padne do PI}) = 95\%$ "chytie medvedia"

$P(\text{odhad IS trafi skutočnú hodnotu } E(Y))$ "chytie matky"



Scheffého simultáne intervaly pre \hat{Y}
predikčné

$$\textcircled{1} \quad \hat{a}^T \hat{\beta} \pm \sqrt{2 F_{2,n-2}(5\%)} \cdot S \sqrt{1 + \hat{a}^T (\hat{X}^T \hat{X})^{-1} \hat{a}}$$



Zajtra bude $x = 70 - 80^\circ\text{F}$... Aký bude ozón?

Kedy použiť ktorý zo zvercov? Treba to mať na vedomie.

1. IS ak na ňa zavíma E / dôsledky prever / parameter

PI ak na ňa zavíma konkrétna hodnota

2. Ak potrebujeme 1 interval $\rightarrow 1/3$, Ak potrebujeme viac intervalov naráz \rightarrow Scheffé

Polyomická regresia

ventilation - ako pumpujú pláca

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2)$$

oxygen - kolko hyslika z nero vedia vytiahnuť

vant. $\hat{\beta}_0 = 3,427$ $\hat{\beta}_1 = -0,01344$

$\hat{\beta}_2 = 0,000008902$

$$a = \begin{pmatrix} 1 \\ X \\ X^2 \end{pmatrix}$$

je m.

Vo zvercoch 1-4 sú použité parametre β

Summary:

call - aktuálny prikazom vznikol objekt

residuals - zvislé odchyly

Coefficients:

Std error - menovateľ T

t value - test. statistika T

p-value pre T:

$$2 \cdot 10^{-16} < 5\% \Rightarrow H_0 \text{ zamietane}$$

$$\frac{\hat{a}^T \hat{\beta} - a^T \hat{\beta}}{S \sqrt{a^T (\hat{X}^T \hat{X})^{-1} a}} \sim t_{n-3} \rightarrow \text{TEST STAT}$$

$$\beta_2 = \begin{pmatrix} 0,0,1 \end{pmatrix} \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix}$$

contrast

$$T = \frac{\hat{\beta}_2 - \beta_2}{S \sqrt{\hat{a}^T (\hat{X}^T \hat{X})^{-1} \hat{a}}} \rightarrow$$

summary
std error

1. $\hat{\beta}_2$ je malý, lebo X -y sú veľké (1000) $\rightarrow \chi^2$ je $10^6 \rightarrow$ ale β_2 je dôležitý

2. $\hat{\beta}_2$ je potrebný, lebo dorázilo

Test. stat., ie $H_0: \beta_0 = 24$ vs. $H_1: \beta_0 \neq 24$ už vedeli sme zo summary \Rightarrow riešenie

summary: residual standard error = S, \$sigma

BDU - používa len a summary otestovať $H_0: \beta_0 = 24$ vs. $H_1: \beta_0 \neq 24$

Poreclán - výsk. ústan závračský cestou do Rače

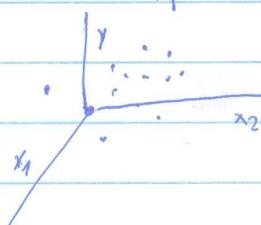
- odolnosť porc. voči vysokym teplotám - teplotné sôky
- ako teplotné sôky uplyvajú na prudie

teplota - 1. sôk, potom chladenie, potom 2. sôk
potom zmenia prudie dôsledky

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

$\hat{\beta}_0 = 2,62$
 $\hat{\beta}_1 = 0,05776$
 $\hat{\beta}_2 = 0,05664$

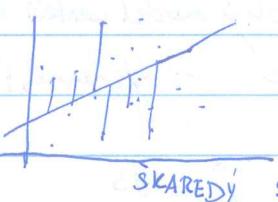
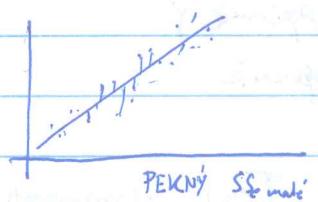
je model dobrý?



12 dát, 12 guliciek + prekladame cerne rovinu

3D obrázok by sme vedeli nárešť, 4D si už nedocenia ani predstavovať

↪ chceme zistíť, či je model dobrý "pomocou číselka"



"krásnú" odmeranie pomocou SSE_e

$SSE_e = \text{miara nehnuteľnosti modelu}$

Aby sme SSE_e vedeli porovnať, vytvoríme ÚBOHÝ MODEL (null model) $Y = \beta_0 + \epsilon$ + zreteľné SSE_e (total) SS_{t}

ÚBOHÝ MODEL JE URČITENÝ KVALITNÝ → $SS_{\text{t}} \geq SSE_e$

a) x_1, x_2 sú dôležité na určenie Y

⇒ úbohy model bude oveľa horší než nás $SS_{\text{t}} > SSE_e$

$$\frac{SSE_e}{SS_{\text{t}}} = 0 \quad \boxed{1 - \frac{SSE_e}{SS_{\text{t}}} = 1}$$

b) x_1, x_2 , nie sú dôležité na určenie Y

⇒ úbohy model nebude oveľa horší než nás $SS_{\text{t}} = SSE_e$

$$\frac{SSE_e}{SS_{\text{t}}} = 1 \quad \boxed{1 - \frac{SSE_e}{SS_{\text{t}}} = 0}$$

Určenie - determinovanie $R^2 = \text{KOEFICIENT DETERMINÁCIE}$ (ako velmi sú x-y potrební na určenie Y)

Summary: multiple R-squared \$r^2\$ squared

Ale: mnohé znaky nehnuteľnosti sa pomocou R^2 zistí nedajú (ale R^2 je užívanejší, lebo Excel)

vedel späťat - R^2

ktorý sôk má výšku uplyn na napätie → ak veríme modelu, je to verenie β_1 a β_2

$$H_0: \beta_1 = \beta_2$$

$$H_1: \beta_1 \neq \beta_2$$

MZR: H_0 zamietame ak $\hat{\beta}_1 \gg \hat{\beta}_2$

$$\left. \begin{array}{l} \hat{\beta}_1 - \hat{\beta}_2 > 0 \\ (0, 1, -1)^T \gg 0 \end{array} \right\} \text{použijeme test pre kontrofakt}$$

H_0 zamietame ak $T \gg 0$

$$T = \frac{a^T \hat{\beta} - 0}{\sqrt{a^T (X^T X)^{-1} a}}$$

H_0 sú nezamietli → test povedal, že neveríme povedali, že β_1 je dôležitosť

↪ malo dát, pomerne malý rozdiel

Test významnosti regresie

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon \rightarrow SSE$$

$$H_0: \beta_1 = 0 \wedge \beta_2 = 0 \quad \text{vs.} \quad H_1: \beta_1 \neq 0 \vee \beta_2 \neq 0$$

$\rightarrow H_1$ hovorí, že úbohy' nestaci'

$$\text{Úbohy' model} \quad Y = \beta_0 + \varepsilon$$

H_0 hovorí, že úbohy' model stačí na popisanie Y

$\rightarrow SST$

$$H_0 \text{ zamietane ak: } SST > SSE$$

$$F = \frac{\frac{SST - SSE}{2}}{\frac{SSE}{12-3}} \left\{ \begin{array}{l} \text{# parametrov zabitých} \\ \frac{SSE}{n-3} = S^2 \end{array} \right.$$

Ak plati H_0 , tak

$$F \sim F_{2, 12-3}$$

p-value $\approx 0 < 5\% \Rightarrow H_0$ zamietane: úbohy' model nestaci' na popisanie Y

summary: F -statistic = test. stat + p-value — významnosť regresie

$$Y = \alpha_0 + \alpha_1 d_1 + \alpha_2 d_2 + \varepsilon$$

napäťe ↑ trvanie 1. ľeden ↑ trvanie 2. ľeden

$$\hat{\alpha}_0 = 93$$

$$\hat{\alpha}_1 = 0,01992$$

$$\hat{\alpha}_2 = 0,02227$$

Test významnosti

$$H_0: \alpha_1 = 0 \wedge \alpha_2 = 0 \quad \text{vs.} \quad H_1: \text{Hovej}$$

$$T = 0,95, \quad p\text{-value} = 41\%$$

Testovanie hypotezy o submodeli

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 \Rightarrow SSE_{\text{model}}$$

ozneme ↑ radiation ↑ temp ↑ wind

$\Rightarrow H_0$ nezamietane

$$R^2 = 0,17 \ll 1$$

$$\hat{\beta}_0 = -0,2978$$

$$H_0: \beta_0 = 0 \wedge \beta_1 = 0 \quad \text{vs.} \quad H_1: \beta_0 \neq 0 \vee \beta_1 \neq 0$$

$$\hat{\beta}_1 = 0,002200$$

submodel stačí

submodel nestaci' na popisanie Y

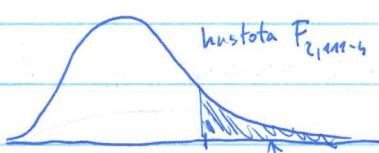
$$\hat{\beta}_2 = 0,05$$

$$\text{Submodel: zohľadňuje } H_0: Y = \beta_2 X_2 + \beta_3 X_3 + \varepsilon \Rightarrow SSE_{\text{submodel}}$$

$$\hat{\beta}_3 = -0,676$$

MZR: H_0 zamietane ak $SSE_{\text{submodel}} > SSE_{\text{model}}$

$$SSE_{\text{submodel}} - SSE_{\text{model}} > 0$$



$$F = \frac{2}{\frac{SSE_{\text{submodel}} - SSE_{\text{model}}}{SSE_{\text{model}}}} \leftarrow \# zabitých parametrov$$

$$1 - \Phi(8,06) = 0,05\% < 5\% \Rightarrow H_0 \text{ zamietane}$$

- venužene naraz výhodit β_0 aj β_1

*alternatívny postup 1: $H_0': \beta_0 = 0$ vs. $H_1': \beta_0 \neq 0$ (test hypotezy o kontraste $\frac{\alpha^\top \beta - 0}{S\alpha^\top (\alpha^\top \alpha)} \rightarrow 5\% \text{ typ Ia}$)

2: $H_0'': \beta_1 = 0$ vs. $H_1'': \beta_1 \neq 0$ (test. hyp. o kontraste) $\rightarrow 5\% \text{ typ Ia}$

pravidlo: H_0 zamietane ak zamietame H_0' alebo $H_0'' \rightarrow$ chyba I. druhu s $\geq 5\%$

! Nedelíť test na viaceré, ak sa da' otiesovať na 1 testom na hladine 5%, použiť ten

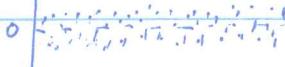
Regresná diagnostika (stredný úvod)

- ako sprojektovať vektorový priestor do \mathbb{R}^2 , aby sme vedeli zistíť, ako dobrá regresia funguje

which=1 : Residuals vs. Fitted values

os X: fitted values = odhadované hodnoty $\hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \hat{\beta}_3 x_{3i}$ (m čísel)

os Y: residuals = $y - (\text{fikované hodnoty})$ (111 čísel)

ideal:  ideally stav = vodorovný smake

- označené dny 20, 23, 77 majú najväčšie rezidua → sú podporné

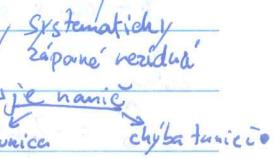
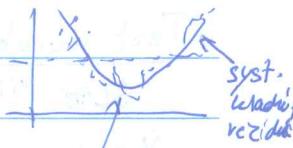
- ak sú obidve cele nad/pod → systematické chyby →

- rozsiahle až do konca

↳ výhľadový vzorový s $\varepsilon_i \sim N(0, \sigma^2)$

↳ σ^2 by mal byť z rovnakého rozdelenia

D(ε_i) je konštantná - HOMOSCEDASTICITY



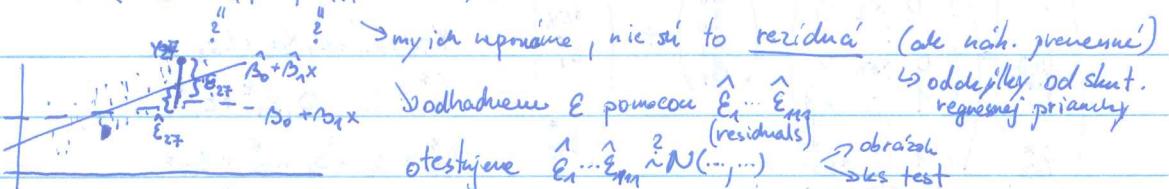
HETEROSCEDASTICITY

↳ teória je celá založená na konštr. → veľký problém

which=2 Scale-location

- absolútne hodnoty rezidui, odmocinene, znormalizujeme, tiež by mal vyzerať ako vodorovný mŕah

which=2 Normal Q-Q - $\varepsilon_1, \dots, \varepsilon_m \stackrel{?}{\sim} N(\dots, \dots)$



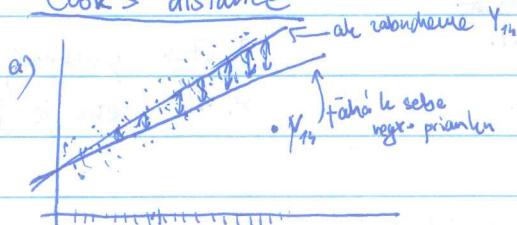
↳ ak zamestnaná normalita? ak ε nemajú norm. rozdelenie

- 1) $\hat{\beta} = \text{LS-ESTIMATOR of } \beta$ - funguje bez ohľadu na rozdelenie ε, len ak sú i.i.d.
- 2) LS / pásy, PI, prediktívne pásy, tesy - X - toto nefunguje, treba použiť nichache' metódy
Neparametrická štatistiká, 5. ročník

toto rozdielnenie neraznivý je 95%

which=4 Cook's distance

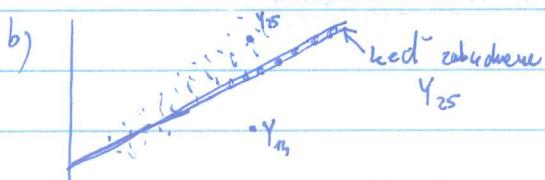
- ako veľmi významné body ovlivňujú regresiu



$$\rightarrow \text{COOK'S DISTANCE}_{14} = \sum_{i=1}^n \left(\frac{Y_i - \hat{Y}_i}{\text{prediktívne ŷ}} \right)^2$$

ak bol Y_{25} outlier → je to veľmi vysoké číslo

COOK'S DISTANCE $_{25} \rightarrow$ je to male číslo



počet cooka sú podľa miest dát 17, 30, 77

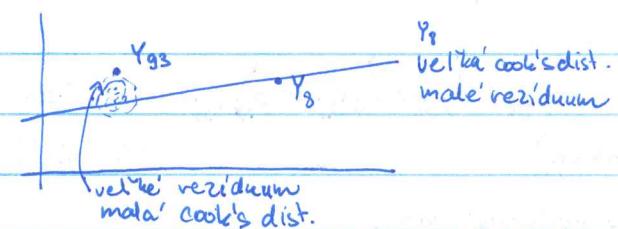
Pre body, ktoré majú vysokú Cook's DIST.
treba skúsiť dátu uvoľniť,

ak sa veľmi zmenia $\hat{\beta}$ (zmena závislosti)

↳ zmena interpretácie

Velké reziduum \Rightarrow Velká cook's distance

Velká cook's dist \Rightarrow Velké reziduum



LEVERAGE POINT (uplyny bod)

\hookrightarrow treba hledat pomocou Cook's distance

Test rovnoběžnosti regresních přímk

SLABÝ VETOR

$$Y_i = \alpha_0 + \alpha_1 X_i + \varepsilon_i \quad (i=1 \dots 53)$$

ozn. \uparrow teplota

$$\hat{\alpha}_0 = -2,69$$

$$\hat{\alpha}_1 = 0,078$$

SILNÝ VETOR

$$Y_i^* = \beta_0 + \beta_1 X_i^* + \varepsilon_i^* \quad (i=1 \dots 53)$$

ozn. \uparrow teplota

$$\hat{\beta}_0 = -0,35$$

$$\hat{\beta}_1 = 0,04$$

Zda sa námže, že slabý vektor znižuje ozn. za rovnalej teploty

$$H_0: \alpha_1 \leq \beta_1 \quad \text{vs.} \quad H_1: \alpha_1 > \beta_1$$

Zložený model

$$\begin{pmatrix} Y_1 \\ \vdots \\ Y_{53} \\ Y_1^* \\ \vdots \\ Y_{53}^* \end{pmatrix} = \begin{pmatrix} 1 & 0 & X_1 & 0 \\ 1 & 0 & X_2 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & X_{53} & 0 \\ 0 & 1 & 0 & X_1^* \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 1 & 0 & X_{53}^* \end{pmatrix} \cdot \begin{pmatrix} \alpha_0 \\ \alpha_1 \\ \vdots \\ \beta_0 \\ \beta_1 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_{53} \\ \varepsilon_1^* \\ \vdots \\ \varepsilon_{53}^* \end{pmatrix}$$

$Y = \underbrace{X}_{\text{ozn.}} \cdot \underbrace{\hat{\alpha}}_{\text{vektor}} +$

$$Y_1 = \alpha_0 + \alpha_1 X_1 + \varepsilon_1$$

$$Y_{53} = \alpha_0 + \alpha_1 X_{53} + \varepsilon_{53}$$

$$Y_1^* = \beta_0 + \beta_1 X_1^* + \varepsilon_1^*$$

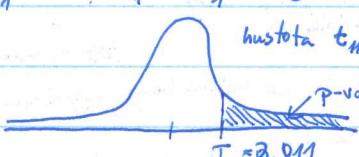
$$Y_{53}^* = \beta_0 + \beta_1 X_{53}^* + \varepsilon_{53}^*$$

$$\text{výdej } \hat{\mu} = \begin{pmatrix} \hat{\alpha}_0 \\ \hat{\alpha}_1 \end{pmatrix}$$

Máme 1 model, můžeme robit test

$$H_0: \alpha_1 \leq \beta_1 \Leftrightarrow \alpha_1 - \beta_1 \leq 0 \Rightarrow \text{test o kontraste } \alpha = (0, 0, 1, 1, -1)$$

$$T = \frac{\hat{\alpha}_1 - \beta_1}{S_{\text{tot}}(X^T X)^{-1} \alpha}$$



$$\text{p-value} = 0,1\% < 5\% \Rightarrow H_0 \text{ zamietane}$$

Je stat. významné, že

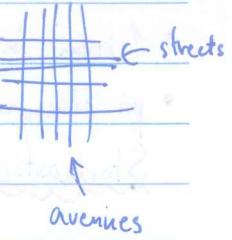
H_0 zamietane ak $\alpha_1 - \beta_1 > 0$

při základu vektora je jiný ozn.

12. prednáška: metoda ANOVA (Fanova) Viacrozmerné statistické analýzy I

Cluster analysis

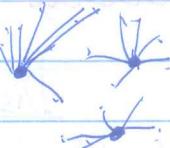
krajiny: kolko bielkovin ziskavajú z rôznych potravín
25 štátov (vektarov) a 9 potravín (priestor R^9)



PAM Partitioning Around Medoids

- určíme počet skupín
bájde medoidy z $\binom{25}{3}$ možnosti
(niektorí z báju)

↳ skúša len tie, ktoré niesú na báji → musia byť med.



→ zmena metriky: Manhattan

→ teraz o podobnosti štátov rozchádzajú veličiny ktoré sú veľké a veľmi kolísané

väčšina	obdobie
1.5	24
0.3	18

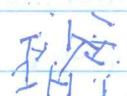
→ normalizácia: $-e \cdot \bar{A}^T$

→ zobrazenie pomocou PCA - Principal Component Analysis - projekcia $R^9 \rightarrow R^2$

Viacrozmerné statistické analýzy 2 (Harman) - Cluster, PCA + Knižka: Kaufmann & Rousseeuw (1990)

LARGE
CLARA - rýchlejšia implementácia k-means clustering

FANNY - pre \neq bod príslušnosti k clusterom (0,1)



FUZZY

AGNES - Agglomerative Nesting - binárne zlukovanie, obvorte sa volá dendrogram

DIVISIVE

- vzd. 2 clusterov min/max/avg

DIANA - opäce ako AGNES → rozdeli jene clustre na 2



- hľadá sa člen clustra, ktorý je najvzdialenejší od $\neq + 1$ reprezentant (prededa)
- rozdelí sa na tých bližších k odiducovi/k reprezentantom

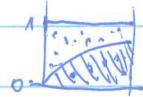
MONA - delenie dát typu yes/no zvierat

jednoduché! vie chodiť? mať košti?

DAISY - Dissimilarity matrix matica 25×25 - vzdialenosť medzi dvojicami \rightarrow upravujeme

Rejection method (Monte-Carlo) integrály

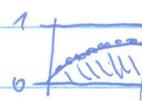
$$\text{Rejection method } \int_0^1 \sin(x) dx = \left[-\cos x \right]_0^1 = 1 - \cos 1 = 0,4596977$$



$$P(\text{traf}_i) = \frac{\text{Splotky}}{S_{\text{vz}}} \Rightarrow \text{Splotky} = S(\text{traf}_i) \cdot S_{\text{vz}}$$

LLN $\frac{\# \text{zásahov}}{\# \text{strieľ}}$

Monte carlo $\int_0^1 \sin(x) dx = \int_0^1 \sin(x) \cdot 1 dx = E(\sin(x)) \stackrel{n}{=} \frac{\sin(x_1) + \sin(x_2) + \dots + \sin(x_n)}{n}$



$$X \dots \rightarrow \text{np. } \rightarrow E(\sin(x)) = \int \sin(x) \cdot f(x) dx$$

- 1 rozm. integrál takto spočítané obdĺžníkmi:
- pri viacrozmerných funkciách: monte-carlo je výhodnejšie
- Stochastické simulácie metódy (Harman) - ^{simulácie} generovanie dejov → viene zistovať vpliv parametrov

Discriminant analysis

Podľa výdychového vzduchu zistovať rizikoviny plúc ^{acetón izopropén chvoj' yes/no} ^{viacrozm. analýza 2}

LDA - Lineárna Discriminácia Analyza (Fisher) ^{priamky oddelené možnosť}  → viene regulovať čístočku vstupu, nie je to úplná lin. separácia + aj viacero priestorov - dokáže oddeľovať len priamkou/parabolou

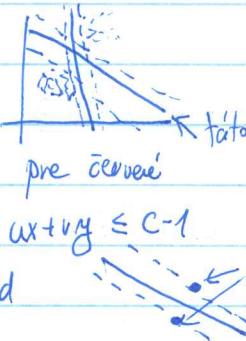
SVM Support vector machine

maximalizuj $\frac{2}{\|w\|^2}$

za podmienky: pre model body

$$wx+wy \geq c+1$$

support vector: podporný bod



$$wx+wy = c+1 \text{ horná}$$

$$wx+wy = c$$

$$wx+wy = c-1 \text{ dolná}$$

$$\begin{aligned} & \text{vzd. priamky} \\ & \frac{|c+1 - (c-1)|}{\sqrt{w^2 + v^2}} \end{aligned}$$

táto priamka je lepšia: väčší počet oholiajúcich

$$wx+wy \leq c-1 \rightarrow \text{kvadratické (rekurzívne) programovanie}$$

support vectors

Kernel trick (Vapnik 1992)

Ak niesú dátá lineárne separovateľné, presunieme ich do vysokozmn. priestoru

a tam sú lineárne sep. \rightarrow SVM

→ prevedomú funkciu Φ netreba poznáť
staci' skalárne súčiny / vlastnosti f .

SVM & soft margins (nie úplne lin. sep.) Cortes & Vapnik 1995

Stochastické optimizačné metódy (Harman) - simulácie ťahanie, evolučné algoritmy